OPINION

A candidate gene approach to searching for low-penetrance breast and prostate cancer genes

*The National Cancer Institute Breast and Prostate Cancer Cohort Consortium**

Abstract | Most cases of breast and prostate cancer are not associated with mutations in known high-penetrance genes, indicating the involvement of multiple low-penetrance risk alleles. Studies that have attempted to identify these genes have met with limited success. The National Cancer Institute Breast and Prostate Cancer Cohort Consortium — a pooled analysis of multiple large cohort studies with a total of more than 5,000 cases of breast cancer and 8,000 cases of prostate cancer — was therefore initiated. The goal of this consortium is to characterize variations in approximately 50 genes that mediate two pathways that are associated with these cancers — the steroid-hormone metabolism pathway and the insulin-like growth factor signalling pathway — and to associate these variations with cancer risk.

With the completion of the Human Genome Project, one of the biggest challenges in genetic research is to identify inherited genetic variants that alter susceptibility to multifactorial and polygenic diseases such as cancer. In the largest available analysis of cancer occurrence among twins, the three types of cancer with the highest estimated heritable components were prostate, colorectal and breast cancer[1]. The polygenic model for cancer susceptibility presumes that the combined effects of variants in many genes, each conferring a small to modest increase in cancer risk, cumulatively accounts for a substantial fraction of this heritable component of risk[2]. These genetic variants might function through interactions with different genes and with behavioural, environmental, or other external risk factors.

New approaches are needed to identify common alleles that confer modest susceptibility to cancer, as opposed to high-penetrance (BOX 1) alleles of genes such as *BRCA1* (breast cancer 1) and *TP53*. Only a small proportion of the estimated number of functional genetic variants has been identified, and the contributions of variants that are located in regulatory, non-coding regions of genes, rather than in exons, are probably also underestimated. Also, much larger studies than those that have been undertaken so far are needed to provide sufficient statistical power to assess small associations, especially those that involve interactions among several genetic variants or between genetic and environmental factors. In the absence of analyses based on larger studies or consortia of studies, the separate publication of numerous small, often contradictory studies continue to generate confusion among researchers, physicians and patients about the genetic and environmental factors that contribute to cancer.

To address this problem, the US National Cancer Institute (NCI) initiated the Cohort Consortium, which will bring together nine cohort studies that are already underway. The goals of this consortium are to assess the effects of common genetic variants in cancer risk and to assess how such variants might interact with each other as well as with lifestyle and environmental risk factors. This would allow researchers to integrate rapid advances in our knowledge of genetic variation into pre-existing large-scale epidemiological studies.

Besides increasing statistical power, there are other advantages to pooling data from multiple cohorts. First, this consortium approach includes only prospective cohort studies, meaning that lifestyle, anthropometric (BOX 1) and environmental risk factor data are collected before patients are diagnosed with cancer. This means that the quality, availability and reliability of this information is not influenced by the presence or absence of disease. The research conducted within the consortium also builds on existing epidemiological resources, avoiding the delays and redundant expense of beginning an entirely new study that would require 10 or more years for cancer cases to accumulate and the results to become available. Another advantage of combining studies in this manner is that participants are drawn from the same defined and enumerated underlying populations — their selection is not influenced by referral patterns, which is a common limitation of hospital-based studies. As common protocols for genetic and epidemiological analyses are used for all of the studies included in the consortium, sources of variation that might afflict meta-analyses of separate published studies can be minimized.

So, what genes should be analysed first in the consortium approach? There is extensive evidence that steroid hormones modulate the risk of developing either breast or prostate cancer, the most common cancers in industrialized nations. Evidence also implicates the insulin-like growth factor (IGF) signalling pathway, which regulates

Box 1 | **Definition of terms used in human genetic analysis**

High-penetrance allele
An allele that is associated with a particular phenotype in a high percentage of carriers.

Single-nucleotide polymorphism
Single-nucleotide polymorphisms (SNPs) occur when a single nucleotide in the DNA sequence is altered.

Anthropometric
Relating to the size, weight or proportions of the human body.

Haplotype
A sequential set of alleles that are present on the same chromosome.

Linkage disequilibrium
The association of a particular allele, such as a genetic variation or mutation associated with a particular disease, with another at a nearby locus more often than would be expected by chance.

Candidate gene
A gene studied for its possible role in a particular trait or disease.

Purifying selection
The loss of an allele from a population, owing to deleterious reductions in reproductive fitness.

Expectation-maximization algorithm
A general method for computing maximum likelihood estimates with incomplete data.

Haplotype imputation
A technique for estimating the number of common haplotypes in a population, often by using the expectation-maximization algorithm, for data in which many SNPs have been genotyped in a set of individuals.

Minor allele frequency
The proportion of alleles at a genetic locus that are of one type in a specific population.

Case–control study
In this study design, cases of disease are identified and their genotypes or past exposure to suspected aetiologic factors are compared with a comparison group of controls who were at risk of developing the disease. The study is said to be nested if the cases and controls are sampled from a predefined ongoing prospective cohort.

Population stratification
A form of confounding by ancestral subgroups if the allele of interest shows marked variation in frequency between subgroups of the population and if these subgroups also differ in their risk of developing the disease.

cellular proliferation and apoptosis[3], in the pathogenesis of these cancers. However, the relative importance of variations in genes that are involved in regulating these pathways, their interaction with environmental factors and their effect on cancer risk is unknown. Furthermore, no previous studies have systematically examined the steroid-hormone metabolism and IGF-signalling pathways in relation to both genotype and phenotype.

So, the first project of the Breast and Prostate Cancer Cohort Consortium (BPC3; BOX 2) is to combine the resources of nine large prospective studies (TABLE 1) and three genomics facilities to characterize the common genetic variations in the products of genes that mediate the steroid-hormone metabolism and IGF-signalling pathways in the aetiology of breast and prostate cancer. The BPC3 is considered a 'proof of principle' study, as it is intended to demonstrate the feasibility and value of this multicentre, multidisciplinary collaboration. What are the aims and biological rationale for the study and the genetic, epidemiological and statistical methods that are being used?

Consortium goals

The overall aim of the Cohort Consortium is to establish a network of epidemiologists, human geneticists, and statisticians to undertake collaborative research on the role of inherited genetic variation, gene–gene and gene–environment interactions in cancer pathogenesis. The project also has four specific aims. First, to comprehensively survey candidate genes that are involved in the steroid-hormone metabolism and IGF-signalling pathways and their receptors, to identify potentially functional polymorphisms and to characterize the haplotype structure across the entire locus of each gene. Second, to investigate the most important effects on breast and prostate cancer risk of single-nucleotide polymorphisms (SNPs) (BOX1), other sequence variants and haplotypes in these genes. Third, to assess whether these gene variants interact with lifestyle and anthropometric factors that are associated with breast and prostate cancer risk. Fourth, to investigate the association of circulating steroid-hormone and IGF concentrations with these genetic variants and with cancer risk, using a subset of the studies in which blood samples were collected before cancer diagnosis.

Candidate genes and cancers

How might genes in the steroid-hormone metabolism and IGF-signalling pathways influence cancer risk? Studies have shown that many of the genes that encode factors in these signalling pathways interact with other gene products and/or environmental factors such as obesity, use of exogenous hormones or smoking. Variants in genes that encode for hormone receptors might modify the activity of circulating hormones (for example, oestrogens, androgens or IGF1) at the tissue level. Variation in genes that control steroidogenesis (such as the *CYP17A2* (cytochrome p450, family 17, subfamily A, polypeptide 1, also

Box 2 | **The Breast and Prostate Cancer Cohort Consortium**

The Breast and Prostate Cohort Consortium (BPC3) is an international group of investigators who study the relationships between variants in candidate genes and breast and prostate cancer risk. The following authors are members of the BPC3 who have contributed to the writing of this document and have agreed with its content (members are listed by writing committee, then in alphabetical order with author affiliations detailed in supplementary information S1 (box)): Hunter, D. J., Riboli, E., Haiman, C. A., Albanes, D., Altshuler, D., Chanock, S. J., Hayes, R. B., Henderson, B. E., Kaaks, R., Stram, D. O., Thomas, G., Thun, M. J. (writing committee); Blanché, H., Buring, J. E., Burtt, N. P., Calle, E. E., Cann, H., Canzian, F., Chen, Y. C., Colditz, G. A., Cox, D. G., Dunning, A. M., Feigelson, H. S., Freedman, M. L., Gaziano, J. M., Giovannucci, E., Hankinson, S. E., Hirschhorn, J. N., Hoover, R. N., Key, T., Kolonel, L. N., Kraft, P., Le Marchand, L., Liu, S., Ma, J., Melnick, S., Pharaoh, P., Pike, M. C., Rodriguez, C., Setiawan, V. W., Stampfer, M. J., Trapido, E., Travis, R., Virtamo, J., Wacholder, S., Willett, W. C.

Table 1 | **Cohort studies in the Breast and Prostate Cancer Cohort Consortium**

Cohort study	Year of blood collection	Initial study design	Number of men with blood sample	Number of women with blood sample	Breast cancer*	Prostate cancer*
American Cancer Society						
Cancer Prevention Study-II	1998	Prospective follow-up study	17,411	21,965	503; 503	1,207; 1,209
Harvard University						
Physicians' Health Study	1982	Randomized trial of aspirin and β-carotene	14,916‡	NA	NA	1,118; 1,447
Nurses' Health Study	1989	Prospective follow-up study	NA	32,826‡	1,100; 1,953	NA
Health Professionals Follow-up Study	1993	Prospective follow-up study	18,410‡	NA	NA	707; 701
Womens' Health Study	1993	Randomized trial of aspirin and vitamin E	NA	28,263‡	705; 705	NA
International Agency for Research in Cancer						
European Prospective Investigation into Cancer and Nutrition	1992	Prospective follow-up study	139,207‡	249,327‡	1,719; 2,844	953; 1,320
Universities of Southern California and Hawaii						
Multiethnic Cohort	1996	Prospective follow-up study	Blood collection ongoing	Blood collection ongoing	1,617; 1,962	2,320; 2,399
National Cancer Institute						
Prostate, Lung, Colon, Ovary	1993	Randomized trial of screening	32,338‡	32,339‡	NA	1,306; 1,668
α-Tocopherol, β-carotene	1991	Randomized trial of β-carotene and vitamin E	NA	26,593‡	NA	1,058; 1,058
Total for all studies	NA	NA	222,282	391,313	5,644; 7,967	8,669; 9,802

*Indicates that data are shown as cases; controls. ‡Indicates that blood specimens were collected before diagnosis. NA, not applicable.

known as 17-α-hydroxylase/17,20-lyase polypeptide 1) allele) have been proposed to influence behavioural factors that affect hormone exposure (such as the use of hormone-replacement therapy[4]) or physiological functions (such as age at menarche[5–7]). The rationale for examining the genes that were chosen for analysis in BPC3, and selected examples of possible gene–environment interactions, are discussed briefly below.

Sex steroids and breast cancer. Abundant evidence indicates that ovarian sex steroids, particularly oestrogens and progesterone, have a key role in promoting breast cancer development. Studies *in vitro* have shown that oestrogens increase breast cell proliferation, and animal studies have shown increased rates of tumour development when the animals are given oestrogens[8]. Prospective cohort studies have shown increases in breast cancer risk in postmenopausal women who have low blood concentrations of sex-hormone-binding globulin (SHBG), and in those who have increased oestrone concentrations, and total and bioavailable oestradiol concentrations[9–12]. Breast cancer risk in postmenopausal women is increased by exposure to exogenous oestrogens and is increased further in women with combined exposure to oestrogens and progestogens[13–21].

Studies also indicate that androgens might affect breast cancer risk. Testosterone and Δ4 androstenedione (Δ4A) are direct precursors of oestrogen synthesis in the ovaries of premenopausal women. After the menopause, plasma androgen concentrations, especially Δ4A, are a key determinant of the amount of oestrogen that is formed by adipose tissue, and circulating concentrations of these hormones are increased in women with breast cancer[14,22–24]. SHBG is produced in the liver, and binds and transports some of the most biologically important oestrogens and androgens in the blood.

Local metabolism of sex steroids in breast tissue also affects oestrogenic activity (FIG. 1). Many steroid-metabolizing enzymes — for example, aromatase (CYP19), which converts Δ4A and testosterone into oestrone and oestradiol, respectively — are active in breast tissue[25]. In addition to the mitogenic properties of oestrogen, the hydroxylation of oestrogens yields metabolites that might be genotoxic[26,27]. The hydroxylation of oestrone and oestradiol occurs through two important pathways: one that involves 16α-hydroxylation to 16α-(OH)oestrone (which is thought to have more potent oestrogenic activity) and oestriol, and the second that leads to catechol (2-(OH) and 4-(OH)) oestrogens[26,28–30]. Because these intermediate metabolites have short half-lives and are difficult to measure *in vivo*, assessing the association(s) of breast cancer risk with functional genetic variants that alter metabolite concentrations might be the most direct approach possible for assessing

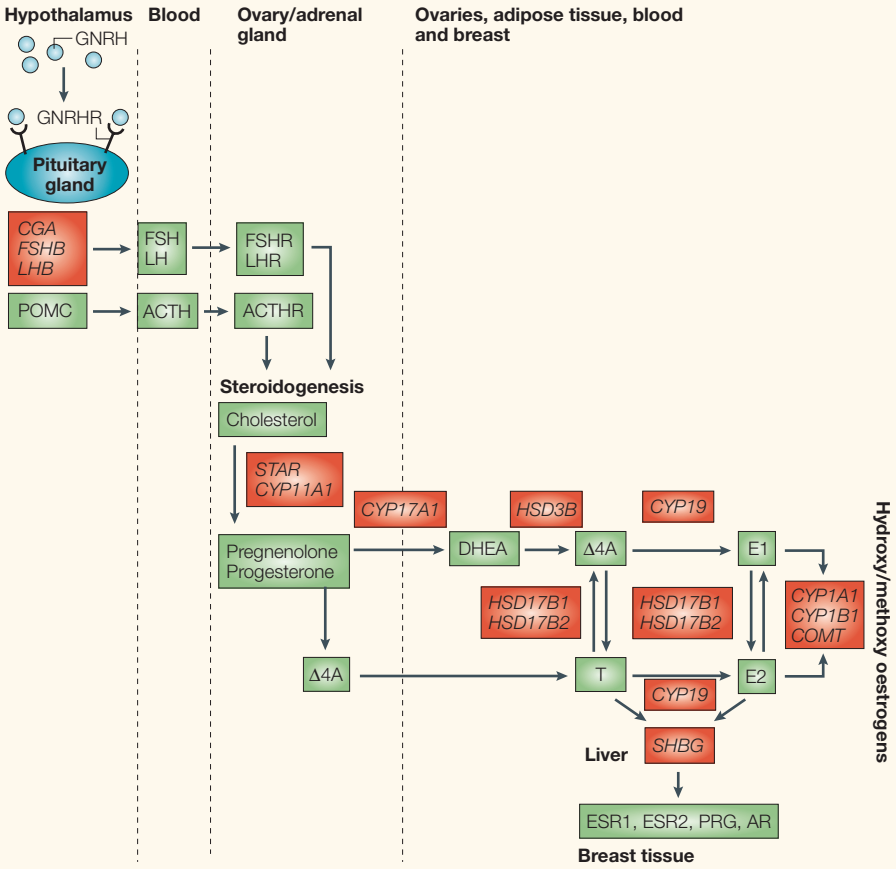

Figure 1 | **Metabolism of sex steroids associated with breast cancer.** Sex-steroid-hormone metabolism is initially regulated by the hypothalamic hormone gonadotropin-releasing hormone (GNRH) through its receptor (GNRHR) to stimulate the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) from the pituitary gland into the bloodstream. FSH and LH share a common α-subunit, encoded by *CGA* (chorionic gonadotropin α-chain), and different β-chains (encoded by *FSHB* (follicle stimulating hormone β-subunit) and *LHB* (luteinizing hormone β-subunit), respectively). These gonadotropins in turn act on their receptors (FSHR and LHR, respectively) in the ovary and adrenal gland to initiate steroid-hormone production. Cholesterol is metabolized through a series of enzymatic steps involving STAR (steroid acute regulatory protein), CYP11A1 (cytochrome p450, family 11, subfamily A, polypeptide 1), CYP17, HSD3B (3-β-hydroxysteroid dehydrogenase, type I) and HSD17B to intermediates such as pregnenolone, progesterone, Δ4 androstenedione (Δ4A) and dehydroepiandrosterone (DHEA, to testosterone (T), oestrone (E1) and oestradiol (E2). Oestrone can also be metabolized from Δ4A in adipose tissue, the most important site of oestrogen production in postmenopausal women. Circulating hormones can be bound to sex-hormone-binding globulin (SHBG) and ultimately exert their activity by binding to the nuclear receptors in cells of the breast tissue, notably the oestrogen receptors (ESR1 and ESR2), the progesterone receptor (PRG) and the androgen receptor (AR). Oestrone and oestradiol can also be further metabolized by enzymes including CYP1A1, CYP1B1 and catechol-*O*-methyltransferase (COMT); these metabolites have been shown to have different potencies in various systems. Adrenocorticotropic hormone (ACTH) is derived from pituitary propiomelanocortin (POMC) and regulates adrenal glucocorticoid production through its receptor ACTHR.

the hypothesis that oestrone and oestradiol metabolites alter this risk.

Gene products that are central to the synthesis, local conversion and hydroxylation/methylation of sex steroids in women, or genes that encode steroid-binding proteins and receptors, are listed in FIG. 1. The BPC3 will take a comprehensive approach to the characterization of variants in these genes and their association with breast cancer risk in a large sample of breast cancer cases and controls.

Sex steroids and prostate cancer. The predominant explanation for the mechanism by which hormones might cause prostate cancer is based on the tumorigenic activities of testosterone and its more potent derivative, dihydrotestosterone (DHT). Prostate development and growth requires androgens, and production of these hormones is inhibited in eunuchs and in men with a constitutional deficiency of 5-α-reductase type II (REFS 28,31), the enzyme that reduces testosterone to DHT. Surgical castration or medical androgen blockade often significantly improves the clinical course of patients with advanced metastatic prostate cancer[32].

Japanese and Chinese migrants to the USA have lower incidence rates of prostate cancer than men of African-American or European ancestry, and also have lower 5-α-reductase activity[33,34]. However, the association between androgen concentrations and prostate cancer has been inconsistent in prospective studies of total or bioavailable testosterone concentrations[35], although small increases in risk have been observed in men who have higher concentrations of andostanediol glucuronide — a serum marker of intraprostatic synthesis of DHT[36]. The lack of consistent evidence linking androgen concentrations with prostate cancer risk could be due to changes in an individual's hormone concentrations over time. FIG. 2 indicates gene products that are related to the initial synthesis and further conversion of sex steroids in men, and might also be associated with prostate cancer risk. The BPC3 will examine variants in these genes with respect to prostate cancer risk in a much larger sample size than available from any of the component studies.

Insulin-like growth factor and breast cancer. There is growing interest in dysregulation of the growth hormone–IGF1 axis as a possible cause of cancer[3,37–39]. IGF1 inhibits apoptosis and stimulates cell proliferation in many cell types, as well as increasing tumour growth[38]. Several case–control studies[40,41] and prospective cohort studies[3,42,43] have shown increases in breast cancer risk in premenopausal women who have higher serum IGF1 concentrations, measured either as absolute concentrations or relative to concentrations of IGF-binding protein 3 (IGFBP3), the principal binding protein of IGF1 in plasma.

Heritability studies have shown that in Western populations, a large part (40–60%) of the variation in IGF1 is (co-) determined by genetic factors[44–46]. Although research to identify genetic determinants of circulating IGF1 and IGFBP3 is intensifying[47–51], no studies have been conducted to search comprehensively for polymorphisms in genes that regulate the synthesis and activity of IGF. No comprehensive attempt has been made to correlate such polymorphisms with inter-subject variations in plasma concentrations of IGF1, IGFBPs and cancer risk. The BPC3 will examine the associations between variants in the IGF1 signalling pathway and plasma concentrations of these proteins in a substantial subset of the

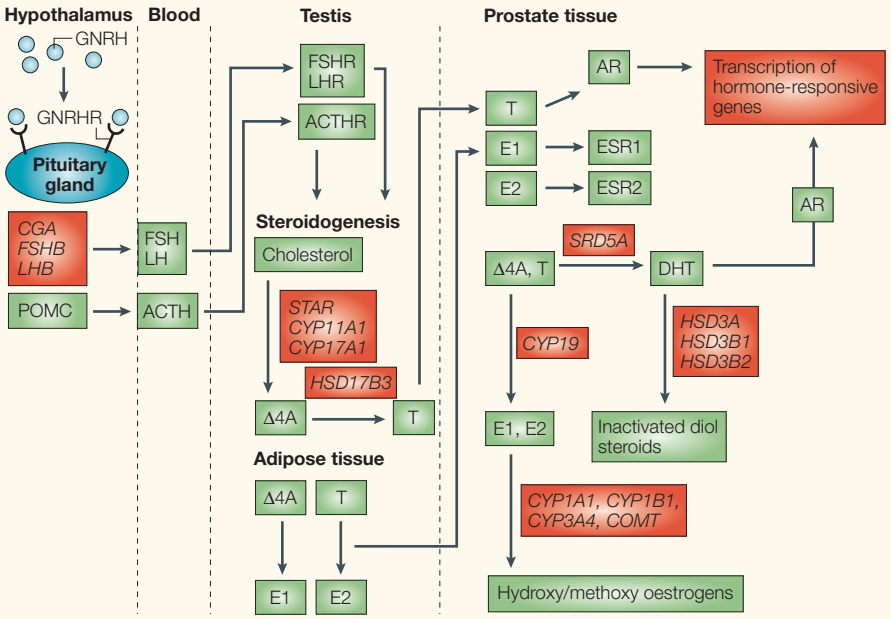

Figure 2 | **Metabolism of sex steroids potentially associated with prostate cancer.** In males, the hypothalamic hormone gonadotropin-releasing hormone (GNRH) binds to its receptor (GNRHR) to stimulate the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) — which share a common α–subunit, encoded by *CGA* (chorionic gonadotropin α-chain) — from the pituitary gland into the bloodstream. When these gonadotropins reach the cells of the testis, they interact with their receptors (FSHR and LHR, respectively) to induce steroid-hormone production. Testosterone (T) is formed in the testis from a similar pathway to that described in FIG. 1, following a series of enzymatic steps that start with cholesterol. The final step is conversion of Δ4 androstenedione (Δ4A) to testosterone by HSD17B3 (17-β-hydroxysteroid dehydrogenase, type III). Testosterone, along with oestrone (E1), oestradiol (E2) and sex-hormone-binding globulin (SHBG, which is produced by the liver), are released into the blood and eventually reach prostate tissue. There, they interact with their receptors (the androgen receptor, AR, and the oestrogen receptors, ESR1 and ESR2) leading to the activation of the enzyme, steroid 5-α-reductase (SRD5A). Δ4A and testosterone are ultimately converted to the active metabolite dihydrotestosterone (DHT), which binds to the androgen receptor with greater affinity than testosterone, leading to the activation of the transcription factor function of the androgen receptor. DHT is inactivated in the prostate by HSD3A, HSD3B1 and HSD3B2. In the prostate, Δ4A and testosterone can also be converted by aromatase (CYP19) into oestrone and oestradiol, which are then further metabolized by CYP1A1 (cytochrome p450, family 1, subfamily A, polypeptide 1), CYP1B1, CYP3A4 and catechol-*O*-methyltransferase (COMT) into hydroxy and methoxy oestrogens. Adrenocorticotropic hormone (ACTH) is derived from pituitary propiomelanocortin (POMC) and regulates adrenal glucocorticoid production through its receptor ACTHR

available cases and controls. FIG. 3 shows gene products involved in the the IGF signalling pathway that have been associated with breast cancer and prostate cancer.

Insulin-like growth factor and prostate cancer. Case–control[52–54] and prospective cohort studies[3] have shown an increase in prostate cancer risk in men who have higher plasma concentrations of IGF1, measured either as absolute concentrations or relative to concentrations of IGFBP3. Increases in IGF1 concentrations that are associated with increased cancer risk might be the result of increased pituitary growth hormone secretion, and might have strong genetic determinants. As in the breast, IGFBPs have a central role in regulating IGF activity in the prostate, and genetic polymorphisms that affect gene expression or protein sequence might therefore also influence prostate cancer risk. The BPC3 will examine these polymorphisms in a very large sample size, and the relation of polymorphisms with IGF plasma concentrations in a substantial subset.

Obesity, endogenous sex steroids and breast cancer. Obesity is a well-established risk factor for breast cancer after the menopause[55]. This relationship is thought to result from an increase in the total and bioavailable plasma oestrogen concentrations, owing to increased peripheral aromatization of androgens[56]. However, obesity can also cause chronic hyperinsulinaemia, and increased insulin decreases hepatic SHBG synthesis. In turn, this might increase ovarian and adrenal production of androgens in some women. Hyperinsulinaemia increases ovarian and adrenal androgen production only in women who are susceptible for ovarian hyperandrogenism (as in women with polycystic ovary syndrome, PCOS)[23], but not in normoandrogenic women[57,58]. This susceptibility is probably genetically determined[59–62].

Candidate genes for breast cancer susceptibility associated with obesity include genes that regulate the synthesis and release of gonadotropins and adrenocorticotropic hormone (*ACTH*), as well as genes that encode enzymes that act in the early steps of gonadal and adrenal steroidogenesis, such as *STAR* (steroid acute regulatory protein), *CYP11A1*, *HSD3B1* (3-β-hydroxysteroid dehydrogenase, type I) and *CYP17A1*. So, one important hypothesis to be examined is whether polymorphisms in these genes, interacting with measures of obesity, are related to breast cancer risk (and at the same time to plasma androgen concentrations). Similar to insulin, IGF1 also stimulates gonadal androgen synthesis[38,39]. Therefore, a relationship between polymorphisms in the IGF1 signalling pathway and steroid-hormone concentrations, perhaps modified by obesity, is plausible. Obesity might also interact with genes that affect oestrogen hydroxylation pathways; obesity has been reported to increase the ratio of 16α-(OH)oestrone to 2α-(OH)oestrone, whereas regular exercise might decrease the ratio[63].

Exogenous hormones. Use of postmenopausal hormones that only contain oestrogen (oestrogen-replacement therapy, ERT) or oestrogen plus progestogen (hormone-replacement therapy, HRT), and, to a lesser extent, use of oral contraceptives, are all associated with some increase in breast cancer risk[12,15,16,17,18,19,20]. Genetic polymorphisms that affect endogenous steroid synthesis might alter circulating hormone concentrations, and might therefore alter menopausal symptoms and influence a woman's decision to use ERT and/or HRT[4]. This could potentially modify the associations of ERT/HRT with breast cancer risk, as has been noted in the case of *CYP17A1* and endometrial cancer[64,65]. The effects of exogenous hormones on breast cancer risk might also be modified by polymorphisms in steroid receptors if these polymorphisms affect the expression or function of receptors. Information on HRT use in the BPC3 studies will be used to examine the relationship between genetic polymorphisms and HRT use, and the potential effect modifications of the relationship between HRT use and breast cancer by these polymorphisms, in the largest sample size available to date.

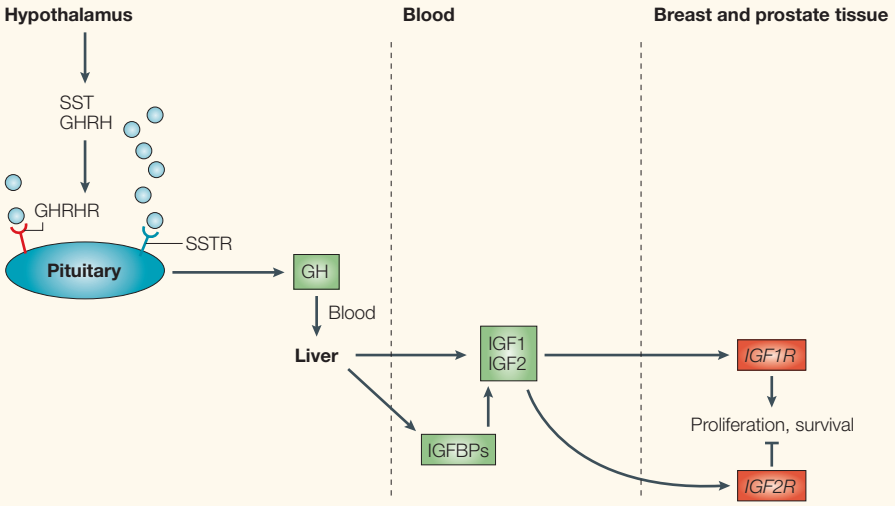

Figure 3 | **The insulin-like growth factor signalling pathway and cell proliferation and survival.** The hypothalamic factors somatostatin (SST) and growth-hormone-releasing hormone (GHRH) control growth hormone (GH) production in the pituitary gland. When growth hormone is released from the pituitary gland into the blood it activates the production of insulin-like growth factor 1 (IGF1) by the liver. IGF1 and IGF2 are then also released into the blood, where they interact with IGF-binding proteins (IGFBPs), which are also produced by the liver. Although there are a number of IGFBPs (IGFBP1–IGFBP6), IGFBP3 provides most of the IGF-binding capacity in the circulation. IGF1 and IGF2 are both ligands for IGF1R, a cell-surface tyrosine kinase receptor that signals proliferative and anti-apoptotic pathways in breast and prostate cells. IGF2R, which binds IGF2, is not known to have tyrosine-kinase activity and might negatively regulate proliferation by reducing the binding of IGF2 to IGF1R. IGF1 reguates sex-hormone-binding globulin (SHBG) synthesis *in vitro*, and plasma IGF1 concentrations have an inverse correlation with plasma SHBG concentrations in both women and men. Furthermore, IGF1 stimulates steroid synthesis in ovarian and testicular tissue[80,81]. Increased IGF1 might therefore affect cancer risk not only through the actions of IGFR1 signalling, but also through an increase in the concentrations of total and bioavailable sex steroids. GHRHR, growth-hormone-releasing hormone receptor; SSTR, somatostatin receptor.

Box 3 | **Common genetic variants and cancer risk**

An underlying hypothesis in the Breast and Prostate Cancer Consortium Cohort activities is that common (defined as allele frequency > 5%) variation in genes accounts for some proportion of inherited susceptibility to cancer. This hypothesis relies on two considerations. The first consideration is based on evolutionary selection. Given the large preponderance of neutral genetic variation that is explained by common variants, only in the setting of purifying selection (BOX 1) against disease-causing alleles will common variants be excluded. Although purifying selection is important for early-onset, severe diseases (such as single-gene, 'monogenic' disorders), it is unclear whether the variants that influence common, late-onset diseases were counter-selected by purifying selection, neutral, or positively selected for during human evolution[91,92]. For sex steroid hormones, the reproductive or developmental impact of one or more functional polymorphisms might be of greater evolutionary relevance in reproductive outcomes (for example, premature birth) than for its effect on late-onset reproductive cancers.

The second consideration is the success in associating common variants with the risk of developing other common diseases, providing a precedent for the relevance of common variant association studies in breast and prostate cancers. There is a growing number of common genetic risk factors that have been reproducibly associated with common diseases, such as *APOE* and Alzheimer disease, *F5* (coagulation factor 5) and thrombosis, four genes (that is, the *HLA* (human luekocyte antigen, also known as major histocompatability complex, *MHC*) locus, *INS* (insulin), *CTLA4* (cytotoxic T-lymphocyte antigen 4) and *PTPN8* (protein tyrosine phosphatase non-receptor, type 8)) and type I diabetes, two genes (*PPARG* (peroxisome proliferative activated receptor-γ) and *KCNJ11* (potassium inwardly-rectifying channel, subfamily J, member 11)) and type II diabetes, *CARD15* (caspase-recruitment-domain family, member 15) and inflammatory bowel disease, *CCR5* (chemokine receptor 5) and HIV infection[84,93], and *CFH* (complement-component factor H) and age-related muscular degeneration[94–96].

Smoking and breast cancer. Smoking decreases the ratio of 16α-(OH)oestrone to 2α-(OH)oestrone[66], but in some studies it has been associated with a modestly increased breast cancer risk[67,68], particularly in early life[69]. The relationships might be modified by polymorphisms in genes that encode the cytochrome p450 enzymes CYP1A1, CYP1B1 or CYP3A4, or other proteins that are involved in carcinogen metabolism[70]. A full examination of the relationship between smoking and breast cancer risk would require a larger number of variants in the many other proteins that are involved in the metabolism of carcinogens in cigarette smoke. The BPC3 will examine the relationships between cigarette smoking and polymorphisms in a subset of carcinogen-metabolizing proteins that are also known to metabolize steroid hormones.

Population genetics. How will the BPC3 identify germline variations in these genes (for a complete list see supplementary information S2 (box)), which might be associated with increased cancer risk? This strategy is based on theoretical and empirical considerations to investigate the role of common genetic variants (BOX 3) associated with sporadic breast and prostate cancer, which are common but complex diseases.

In surveying each gene of interest, there are two parallel approaches for identifying genetic variants that might increase cancer risk (BOX 4). The first is to identify SNPs that occur more often in patients with breast or prostate cancer. In the BPC3 study, this requires systematic resequencing of all exons (and some conserved non-coding regions) in candidate genes from germline DNA samples isolated from 190 patients with advanced breast or prostate cancer (FIG. 4). To indirectly assess associations with putative non-coding changes, the study will use an approach that is based on tagging common haplotypes (BOX1) because functional variants that influence gene regulation (rather than protein sequence) are often found outside of (and sometimes at a considerable distance from) exons. This type of 'haplotype mapping' is important because it does not require direct sequencing of the non-coding regions in each gene locus.

SNPs in the genes to be analysed have also been reported in public databases, and these can also be used in haplotype determination — this data can be merged with that discovered by the sequencing process shown in supplementary information S3 (figure). A set of high frequency SNPs that have a minor allele frequency of > 5% are then selected

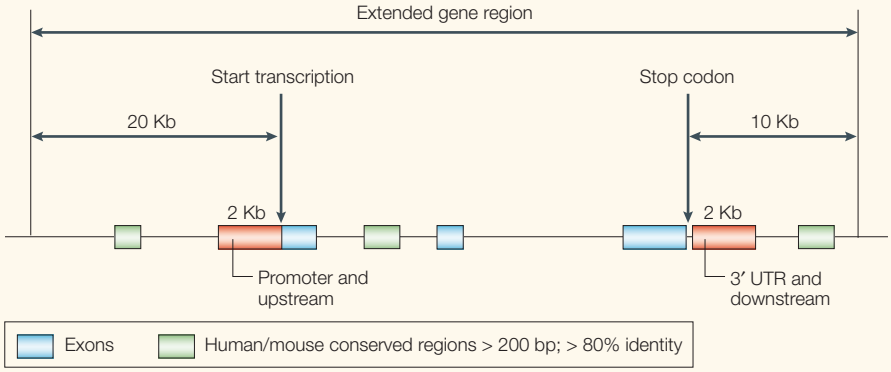

Figure 4 | **SNP identification by data mining and by resequencing the BPC3 SNP discovery panel.** Genetic variations known as single-nucleotide polymorphisms (SNPs) are analysed in candidate cancer genes for their association with cancer. When a gene is selected for SNP analysis, a region starting 20 Kb upstream of the transcription start site and ending 10 Kb downstream of the stop codon is scanned in available databases to include in the analysis all previously identified SNPs with known minor allele frequency > 5% in the European population. This region is called the 'extended gene region'. It can occasionally include parts of adjacent genes. In addition, to identify unknown SNPs, all exons and for some genes the 2 Kb regions upstream of the transcription start site and downstream of the stop codon (the 3′ untranslated region, UTR) are sequenced in a panel of 190 DNAs from patients with cancer. For some genes, additional regions that are at least 200 base pairs long and demonstrate 80% identity with the mouse homologues are also sequenced because they are probably important for gene regulation. The 190 DNA samples included in the SNP discovery panel of the Breast and Prostate Cancer Consortium Cohort (BPC3), are taken from 5 ethnic groups (African-American, Asian-American, European, Hawaiian or Latino) and include, for each group, 19 patients with advanced cases of breast cancer and 19 patients with advanced cases of prostate cancer. This sample size was chosen to provide > 85% power to discover variants with a minor allele frequency > 5% that is present in a single population group. All SNPs that occur more than once in the study population are included in the subsequent SNP selection process (supplementary information S4 (figure)). Furthermore, non-redundant, non-synonymous (missense and insertion/deletion muations) SNPs that were observed more than once in a given population were genotyped in the corresponding case–control groups (supplementary information S4 (figure)).

and genotyped in the haplotyping panel whenever possible. Using the algorithm shown in supplementary information S4 (figure), SNPs to be genotyped in the nested case–control studies will be selected[82,83]. All data from SNP-discovery efforts and haplotype structure (BOX1) characterization will be made publicly available (see the USC–Norris Comprehensive Cancer Center) so that they can be used to guide subsequent research. Sequence-validated SNP assays will also be posted online (see SNP500)[71].

Study populations

So far, the BPC3 study includes samples from 5,347 patients with breast cancer and 8,669 patients with prostate cancer from the existing cohort of over 800,000 men and women. Approximately 610,000 of these study participants have given blood samples (some samples from cases and controls are collected retrospectively in the Multiethnic Cohort (MEC) and American Cancer Society studies). The BPC3 includes DNA samples from all NCI-funded prospective studies that are known to include at least 500 patients who developed breast or prostate cancer by 2002. The key characteristics of the component studies are displayed in TABLE 1 and are described in detail elsewhere[10,72–78]. Participants in these studies live in all 50 US states, and 11 countries in Europe. The MEC (TABLE 1) is the only study involved in the consortium in which a deliberate oversampling of ethnic minorities was conducted; this cohort enrolled substantial numbers of native Hawaiians, Latinos, Japanese-Americans and African-Americans, in addition to Caucasians[21]. More than 90% of participants in all studies are still being followed, and patients who develop cancer can be identified through various mechanisms such as cancer and death registries, and medical records obtained after initial self-reporting.

The datasets consist of a series of case–control studies (BOX 1) nested within these cohorts. Patients with breast and prostate cancer in each cohort were matched in a ratio of at least one control per case by age, ethnicity and, in some cohorts, by other potential confounding factors. In the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort (TABLE 1), ethnic subclusters correspond to regions of recruitment, so cases and controls are also matched by recruitment centre. One advantage to the BPC3 is that results can be examined in each cohort separately and data can then be combined in a pooled analysis to maximise power to associate specific genetic variations with cancer type, other genetic variations, and the effects of exposure to various lifestyle and environmental factors.

How will BPC3 be carried out?

One of the challenges of such a large study is to efficiently genotype a large number of samples. In the BPC3 studies, genotyping would be performed in the cohort-specific laboratories, using common protocols and reagents. Before starting case–control analyses for each SNP, each centre genotypes a set of 94 samples from the diversity panel of lymphoblastoid cell lines from populations around the world maintained by CEPH (Centre d'Etude du Polymorphisme Humain) and used by the SNP500.

Box 4 | **Linkage disequilibrium in the human genome**

A portion of the inherited risk underlying breast and prostate cancer might reside in pathways that have already been implicated in the diseases by epidemiological and pathophysiological evidence. In candidate genes (BOX1) identified in these pathways, some of the genetic variants that contribute to cancer risk could exist at an appreciable frequency in the population, therefore justifying the conduct of association studies of common variants (here defined as those with minor allele frequencies > 5%) and breast and prostate cancers[84]. When 2 randomly chosen copies of the human genome are compared, they typically differ at only 1 in 1,000 nucleotides[85–87]. For heterozygous sites, approximately 90% are owing to variants for which the frequency in the general population is > 1% (REFS 88–90). Nearly all of the commonly occurring variants in current populations derive from a small ancestral population dating back at least 100,000–200,000 years in which low mutation rate and substantial genetic drift has limited genetic diversity. Moreover, because of low recombination rates per generation, genetic variants often display correlations with neighbouring variants (known as linkage disequilibrium) resulting in a limited number of frequent haplotypes. In such regions, it is possible to select a subset of markers that effectively captures the common genetic variation.

Another challenge of the study will be to assay plasma steroid-hormone concentrations of so many samples. The consortium study requires the analysis of concentrations of plasma steroid-hormones and IGFs in approximately 2,500 samples from breast cancer cases and 3,000 samples from prostate cancer cases arising among participants in the EPIC, the Harvard cohorts, α-tocopherol β-carotene, and the Prostate, Lung, Colon, Ovary studies. Plasma hormone concentrations to be analysed in the breast cancer studies include oestradiol, oestrone, testosterone, Δ4A, dehydroepiandrosterone-sulphate (DHEAS), SHBG, IGF1 and IGFBP3. In the prostate cancer study, researchers will assay plasma concentrations of testosterone, Δ4A, SHBG, androstane-17, 3-diol-glucuronide (adiol-G), IGF1 and IGFBP3 from study participants. Most of these samples have also been analysed in a single laboratory at the International Agency for Research on Cancer (IARC), therefore minimizing inter-laboratory measurement differences.

A unique feature of the BPC3 is the opportunity to examine the heterogeneity of associations between the component studies. Even among the primarily Caucasian cohorts, the geographical diversity represented by the BPC3 mitigates the possibility that hidden population stratification (BOX 1) would lead to false-positive results in the same direction in all cohorts simultaneously[79]. To assess homogeneity, analyses include tests for the significance of apparent inconsistencies in the effect of a single SNP or haplotype by calculating their effect for each subcohort or ethnic group of interest. We apply the same approach to the tests of gene–environment interactions that will provide estimates of whether the genetic effects vary according to non-genetic breast and prostate cancer risk factors in a much larger sample size than available to date.

Conclusions

The BPC3 was created to examine, in a coordinated manner, the association of cancer risk with low-penetrance candidate susceptibility genes and their interaction with exogenous or endogenous exposures. Previous consortia that were designed to address the genetic causes of breast and prostate cancer have been involved in the joint analysis of family-based studies, and have therefore been limited to the search for genetic variants with relatively high penetrance. Taking a candidate gene approach and combining association studies presents substantial challenges, mostly because of the low relative risks that are predicted for any single genetic factor or combination of factors, and the previously low probability that any specific variant, among the many actual variants, is causally associated with the disease of interest. The approach used by the BPC3 is designed to use available technologies to examine the 'common-variant–common-disease' hypothesis (BOX 3). Our current approach does not evaluate an alternative hypothesis — that multiple rare variants in a gene are associated with low relative risks — as this hypothesis cannot be examined in a cost-effective manner in large populations, but requires further development of technologies to permit the examination of all known variations in a gene.

The BPC3 was initiated because having multiple groups working independently on specific candidate genes has resulted in inconsistent findings owing to chance, the inadequate sample size of individual studies to evaluate gene–environment interactions, publication bias and the lack of comprehensive characterization of genetic variants of interest. The BPC3 attempts to ameliorate these problems by taking a collaborative approach to examine candidate genes in two pathways that have been previously associated with breast and prostate cancer pathogenesis. The BPC3 combines the resources of existing cohort studies with recent advances in population genetics to conduct adequately powered association studies.

As whole-genome SNP scans identify new candidate genes and regions that might be involved in breast and prostate cancer, the BPC3 will provide a resource for the replication of these findings. The current analysis will be based on a more comprehensive characterization of variants in candidate genes in the steroid-hormone metabolism and IGF-signalling pathways than has been previously available. We expect that examination of both individual SNPs and haplotypes in these genes in relation to breast and prostate cancer will determine which, if any, of the inherited variants are associated with risk, as well as evidence for those that are not.

Members of the BPC3 who have agreed to be named are listed in Box 2.

Correspondence to David Hunter at the Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 677 Huntington avenue, Boston, Massachusetts 02115, USA.
e-mail: dhunter@hsph.harvard.edu

doi:10.1038/nrc1754

1. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
2. Pharoah, P. D., Dunning, A. M., Ponder, B. A. & Easton, D. F. Association studies for finding cancer-susceptibility genetic variants. *Nature Rev. Cancer* **4**, 850–860 (2004).
3. Pollak, M. N., Schernhammer, E. S. & Hankinson, S. E. Insulin-like growth factors and neoplasia. *Nature Rev. Cancer* **4**, 505–518 (2004).
4. Feigelson, H. S. *et al.* Cytochrome *P450c17*α gene (*CYP17*) polymorphism predicts use of hormone replacement therapy. *Cancer Res.* **59**, 3908–3910 (1999).
5. Thompson, P. A. & Ambrosone, C. Molecular epidemiology of genetic polymorphisms in estrogen metabolizing enzymes in human breast cancer. *J. Natl Cancer Inst. Monogr.* **27**, 125–134 (2000).
6. Feigelson, H. S., Coetzee, G. A., Kolenel, L. N., Ross, R. K. & Henderson, B. E. A polymorphism in the *CYP17* gene increases the risk of breast cancer. *Cancer Res.* **57**, 1063–1065 (1997).
7. Haiman, C. A. *et al.* The relationship between a polymorphism in *CYP17* with plasma hormone levels and breast cancer. *Cancer Res.* **59**, 1015–1020 (1999).
8. Henderson, B. E. & Feigelson, H. S. Hormonal carcinogenesis. *Carcinogenesis* **21**, 427–433 (2000).
9. Thomas, H. V. *et al.* A prospective study of endogenous serum hormone concentrations and breast cancer risk in premenopausal women on the island of Guernsey. *Br. J. Cancer* **75**, 1075–1079 (1997).
10. Hankinson, S. E. *et al.* Plasma sex steriod hormone levels and risk of breast cancer in postmenopausal women. *J. Natl Cancer Inst.* **90**, 1292–1299 (1998).
11. Cauley, J. A. *et al.* Elevated serum estradiol and testosterone concentrations are associated with high risk for breast cancer. Study of Osteoporotic Fractures Research Group. *Ann. Intern. Med.* **130**, 270–277 (1999).
12. Key, T., Appleby, P., Barnes, I. & Reeves, G. Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. *J. Natl Cancer Inst.* **94**, 606–616 (2002).
13. Key, T. J. & Pike, M. C. The role of oestrogens and progestogens in the epidemiology and prevention of breast cancer. *Eur. J. Cancer Clin. Oncol.* **24**, 29–43 (1988).
14. Bernstein, L. & Ross, R. Endogenous hormones and breast cancer risk. *Epidemiol Rev.* **15**, 48–65 (1993).
15. Ross, R. K., Paganini-Hill, A., Wan, P. C. & Pike, M. C. Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J. Natl Cancer Inst.* **92**, 328–332 (2000).
16. Olsson, H., Bladstrom, A., Ingvar, C. & Moller, T. R. A population-based cohort study of HRT use and breast cancer in southern Sweden. *Br. J. Cancer* **85**, 674–677 (2001).
17. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53,297 women with breast cancer and 100,239 women without breast cancer from 54 epidemiological studies. *Lancet* **347**, 1713–1727 (1996).
18. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiologic studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* **350**, 1047–1059 (1997).
19. Rossouw, J. E. *et al.* Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* **288**, 321–333 (2002).
20. Anderson, G. L. *et al.* Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA* **291**, 1701–1712 (2004).
21. Kolonel, L. N., Altshuler, D. & Henderson, B. E. The Multiethnic Cohort study: exploring genes, lifestyle and cancer risk. *Nature Rev. Cancer* **4**, 519–527 (2004).
22. Secreto, G. & Zumoff, B. Abnormal production of androgens in women with breast cancer. *Anticancer Res.* **14**, 2113–2117 (1994).
23. Kaaks, R. Nutrition, hormones, and breast cancer: is insulin the missing link? *Cancer Causes Control* **7**, 605–625 (1996).
24. Ehrmann, D. A., Barnes, R. B. & Rosenfield, R. I. Polycystic ovary syndrome as a form of functional ovarian hyperandrogenis due to dysregulation of androgen secretion. *Endocr. Rev.* **16**, 322–353 (1995).
25. Zhu, B. T. & Conney, A. H. Functional role of estrogen metabolism in target cells: review and perspective. *Carcinogenesis* **19**, 1–27 (1998).
26. Jefcoate, C. R. *et al.* Tissue-specific synthesis and oxidative metabolism of estrogens. *J. Natl Cancer Inst. Monogr.* **27**, 95–112 (2000).
27. Yue, W. *et al.* Genotoxic metabolites of estradiol in breast: potential mechanism of estradiol induced carcinogenesis. *J. Steroid Biochem. Mol. Biol.* **86**, 477–486 (2003).
28. Zhu, Y. S., Katz, M. D. & Imperato-McGinley, J. Natural potent androgens: lessons from human genetic models. *Baillieres Clin. Endocrinol. Metab.* **12**, 83–113 (1998).
29. Ball, P. & Knuppen, R. Catecholoestrogens (2-and 4-hydroxyoestrogens): chemistry, biogenesis,

metabolism, occurrence and physiological significance. *Acta Endocrinol. Suppl. (Copenh.)* **232**, 1–127 (1980).
30. Martucci, C. P. & Fishman, J. P450 enzymes of estrogen metabolism. *Pharma. Ther.* **57**, 237–257 (1993).
31. Randall, V. A. Role of 5 α-reductase in health and disease. *Baillieres Clin. Endocrinol. Metab.* **8**, 405–431 (1994).
32. Bosland, M. C. Hormonal factors in carcinogenesis of the prostate and testis in humans and in animal products. *Prog. Clin. Biol. Res.* **394**, 309–352 (1996).
33. Lookingbill, D. P. *et al.* Clinical and biochemical parameters of androgen action in normal healthy Caucasian versus Chinese subjects. *J. Clin. Edocrinol. Metab.* **72**, 1242–1248 (1991).
34. van Houten, M. E. & Gooren, L. J. Differences in reproductive endocrinology between Asian men and Caucasian men—a literature review. *Asian J. Androl.* **2**, 13–20 (2000).
35. Platz, E. A. & Giovannucci, E. The epidemiology of sex steroid hormones and their signaling and metabolic pathways in the etiology of prostate cancer. *J. Steroid Biochem. Mol. Biol.* **92**, 237–253 (2004).
36. Kaaks, R., Lukanova, A. & Sommersberg, B. Plasma androgens, IGF-1, body size, and prostate cancer risk: a synthetic review. *Prostate Cancer Prostatic Dis.* **3**, 157–172 (2000).
37. Yu, H. & Rohan, T. Role of the insulin-like growth factor family in cancer development and progression. *J. Natl. Cancer Inst.* **92**, 1472–1489 (2000).
38. Khandwala, H. M., McCutcheon, I. E., Flyvbjerg, A. & Friend, K. E. The effects of insulin-like growth factors on tumorigenesis and neoplastic growth. *Endocr. Rev.* **21**, 215–244 (2000).
39. Kaaks, R. & Lukanova, A. Energy balance and cancer: the role of insulin and insulin-like growth factor-I. *Proc. Nutr. Soc.* **60**, 91–106 (2001).
40. Peyrat, J. P. *et al*. Plasma insulin-like growth factor-1 (IGF-1) concentrations in human breast cancer. *Eur. J. Cancer* **29A**, 492–497 (1993).
41. Bohlke, K., Cramer, D. W., Trichopoulos, D. & Mantzoros, C. S. Insulin-like growth factor-I in relation to premenopausal ductal carcinoma *in situ* of the breast. *Epidemiology* **9**, 570–573 (1998).
42. Toniolo, P. *et al.* Serum insulin-like growth factor-I and breast cancer. *Int. J. Cancer* **88**, 828–832 (2000).
43. Hankinson, S. E. *et al.* Circulating concentrations of insulin-like growth factor-I and risk of breast cancer. *Lancet* **351**, 1393–1396 (1998).
44. Hall, K., Hilding, A. & Thoren, M. Determinants of circulating insulin-like growth factor-I. *J. Endocrinol Invest.* **22**, 48–57 (1999).
45. Hong, Y., Brismar, K., Hall, K., Pedersen, N. L. & de Faire, U. Associations between insulin-like growth factor-I (IGF-I), IGF-binding protein-1, insuin and other metabolic measures after controlling for genetic influences: results from middle-aged and elderly monozygotic twins. *J. Endocrinol.* **153**, 251–257 (1997).
46. Harrela, M. *et al.* Genetic and environmental components of interindividual variation in circulating levels of IGF-I, IGF-II, IGFBP-1, and IGFBP-3. *J. Clin. Invest.* **98**, 2612–2615 (1996).
47. Lopez-Bermejo, A., Buckway, C. K. & Rosenfeld, R. G. Genetic defects of the growth hormone-insulin-like growth factor axis. *Trends Endocrinol. Metab.* **11**, 39–49 (2000).
48. Jernstrom, H. *et al.* Genetic and nongenetic factors associated with variation of plasma levels of insulin-like growth factor-I and insulin-like growth factor-binding protein-3 in healthy premenopausal women. *Cancer Epidemiol. Biomarkers Prev.* **10**, 377–384 (2001).
49. Rasmussen, S. K. *et al.* Studies of the variability of the genes encoding the insulin-like growth factor I receptor and its ligand in relation to type 2 diabetes mellitus. *J. Clin. Endocrinol. Metab.* **85**, 1606–1610 (2000).
50. Sun, G., Chagnon, M. & Bouchard, C. A common polymorphism in the human insulin-like growth factor binding protein 3 gene. *Mol. Cell. Probes* **14**, 55–56 (2000).
51. Sasi, R., Puebla, L., Khare, S. & Patel, Y. C. Polymorphism in the 5′ flanking region of the human somatostatin receptor subtype 5. *Gene* **214**, 45–49 (1998).
52. Mantzoros, C. S. *et al.* Insulin-like growth factor 1 in relation to prostate cancer and benign prostatic hyperplasia. *Br. J. Cancer* **76**, 1115–1118 (1997).
53. Wolk, A. *et al.* Insulin-like growth factor 1 and prostate cancer risk: a population-based, case-control study. *J. Natl Cancer Inst.* **90**, 911–915 (1998).
54. Chokkalingam, A. P. *et al.* Insulin-like growth factors and prostate cancer: a population-based case-control study in China. *Cancer Epidemiol. Biomarkers. Prev.* **10**, 431–437 (2001).
55. van den Brandt, P. A. *et al.* Pooled analysis of prospective cohort studies on height, weight and breast cancer risk. *Am. J. Epidemiol.* **152**, 514–527 (2000).
56. Siiteri, P. K. Adipose tissue as a source of hormones. *Am. J. Clin. Nutr.* **45**, 277–282 (1987).
57. Austin, H., Austin, J. M., Jr., Partridge, E. E., Hatch, K. D. & Shingleton, H. M. Endometrial cancer, obesity, and body fat distribution. *Cancer Res.* **51**, 568–572 (1991).
58. Newcomb, P. A. *et al.* Association of dietary and life-style factors with sex hormones in postmenopausal women. *Epidemiol.* **6**, 318–321 (1995).
59. Franks, S. Polycystic ovary syndrome. *N. Engl. J. Med.* **333**, 853–861 (1995).
60. Dunaif, A. Insulin resistance and the polycystic ovary syndrome: mechanism and implications for pathogenesis. *Endocr. Rev.* **18**, 774–800 (1997).
61. Urbanek, M. *et al.* Thirty-seven candidate genes for polycystic ovary syndrome: strongest evidence for linkage is with follistatin. *Proc. Natl Acad. Sci. USA* **96**, 8573–8578 (1999).
62. Legro, R. S. *et al.* Phenotype and genotype in polycystic ovary syndrome. *Recent Prog. Hormone Res.* **53**, 217–256 (1998).
63. De Cree, C. *et al.* Response of catecholestrogen metabolism to acute graded exercise in normal menstruating women before and after training. *J. Clin. Endocrinol. Metab.* **82**, 3342–3348 (1997).
64. Haiman, C. A., Hankinson, S. E., Colditz, G. A., Hunter, D. J. & De Vivo, I. A polymorphism in *CYP17* and endometrial cancer risk. *Cancer Res.* **61**, 3955–3960 (2001).
65. McKean-Cowdin, R. *et al.* Risk of endometrial cancer and estrogen replacement therapy history by *CYP17* genotype. *Cancer Res.* **61**, 848–849 (2001).
66. Michnovicz, J. J., Naganuma, H., Hershcopf, R. J., Bradlow, H. L. & Fishman, J. Increased urinary catechol estrogen excretion in female smokers. *Steroids* **52**, 69–83 (1988).
67. Couch, F. J. *et al.* Cigarette smoking increases risk for breast cancer in high-risk breast cancer families. *Cancer Epidemiol. Biomarkers Prev.* **10**, 327–332 (2001).
68. Johnson, K. C., Hu, J. & Mao, Y. Passive and active smoking and breast cancer risk in Canada, 1994–97. The Canadian Cancer Registries Epidemiology Research Group. *Cancer Causes Control* **11**, 211–221 (2000).
69. Terry, P. D. & Rohan, T. E. Cigarette smoking and the risk of breast cancer in women: a review of the literature. *Cancer Epidemiol. Biomarkers Prev.* **11**, 953–971 (2002).
70. Mitrunen, K. & Hirvonen, A. Molecular epidemiology of sporadic breast cancer. The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. *Mutat Res.* **544**, 9–41 (2003).
71. Packer, B. R. *et al.* SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res.* **32**, D528–D532 (2004).
72. Calle, E. E. *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490–2501 (2002).
73. The α-Tocopherol β-Carotene Cancer Prevention Study Group. The α-Tocopherol β-Carotene Lung Cancer Prevention Study: design, methods, participant characteristics, and compliance. *Ann. Epidemiol.* **4**, 1–10 (1994).
74. Riboli, E. *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
75. Hayes, R. B. *et al.* Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clin. Trials* **21**, 349S–355S (2000).
76. Sesso, H. D. *et al.* Comparison of baseline characteristics and mortality experience of participants and nonparticipants in a randomized clinical trial: the Physicians' Health Study. *Control Clin. Trials* **23**, 686–702 (2002).
77. Chan, J. M. *et al.* Insulin-like growth factor-I (IGF-I) and IGF binding protein-3 as predictors of advanced-stage prostate cancer. *J. Natl Cancer Inst.* **94**, 1099–1106 (2002).
78. Giovannucci, E. *et al.* Nutritional predictors of insulin-like growth factor I and their relationships to cancer in men. *Cancer Epidemiol. Biomarkers Prev.* **12**, 84–89 (2003).
79. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
80. Cara, J. F. Insulin-like growth factors, insulin-like growth factor binding proteins and ovarian androgen production. *Horm. Res.* **42**, 49–54 (1994).
81. Lin, T., Haskell, J., Vinson, N. & Terracio, L. Direct stimulatory effects of insulin-like growth factor-I on Leydig cell steroidogenesis in primary culture. *Biochem. Biophys. Res. Commun.* **137**, 950–956 (1986).
82. Stram, D. O. *et al.* Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55**, 27–36 (2003).
83. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
84. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33** (suppl.), 228–237 (2003).
85. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
86. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
87. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes fro blood-pressure homeostatsis. *Nature Genet.* **22**, 239–247 (1999).
88. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
89. Marth, G. *et al.* Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet.* **27**, 371–372 (2001).
90. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
91. Harpending, H. & Rogers, A. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* **1**, 361–385 (2000).
92. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
93. Rosand, J. & Altshuler, D. Human genome sequence variation and the search for genes influencing stroke. *Stroke* **34**, 2512–2516 (2003).
94. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
95. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
96. Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).

Acknowledgements
The authors gratefully acknowledge the participants in the component cohort studies, and the expert contributions of: Hardeep Ranu, Craig Labadie, Lisa Cardinale and Shamika Ketkar from Harvard University, USA; William Modi, Merideth Yeager, Robert Welch, Cynthia Glaser and Laurie Burdett from the National Cancer Institute; Mourad Sahbatou and Emmanuel Tubacher from Centre d'Etude du Polymorphism Humain; and Loreall Pooler, Stephanie Riley, John Casagrande and Reed Comire from the University of Southern California, USA.

Competing interests statement
The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ACTH | *BRCA1* | *CYP17A1* | IGF1 | IGFBP3 | SHBG | *TP53* |
National Cancer Institute: http://www.cancer.gov
breast cancer | prostate cancer

FURTHER INFORMATION

Core genotyping facility website of the NCI: http://cgf.nci.nih.gov
dpSNP: http://www.ncbi.nlm.nih.gov/SNP/index.html
Haploview: http://www.broad.mit.edu/mpg/haploview
Haplotype tagging SNP (htSNP) selection in the Multiethnic Cohort Study: http://www-rcf.usc.edu/~stram/tagSNPs.html
Sequence validated SNP assays (SNP500): http://snp500cancer.nci.nih.gov
USC–Norris Comprehensive Cancer Center: http://www.uscnorris.com/MECGenetics

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box) | S3 (figure) | S4 (figure)
Access to this links box is available online.